QÜESTIIÓ, vol. 22, 2, p. 291-309, 1998

ESTIMACIÓN DEL NÚMERO DE CLUSTERS EN UNA POBLACIÓN APLICANDO EL JACKKNIFE GENERALIZADO

J.J. PRIETO MARTÍNEZ
Universidad Carlos III de Madrid*

Sea una población constituida por un número desconocido de clusters. Este trabajo desarrolla una secuencia finita de estimadores no paramétricos para el número de clusters, basándose en el método jackknife generalizado. Estos estimadores resultan ser una combinación lineal de las frecuencias de observación de cada cluster. Se propone entonces un procedimiento de selección para elegir el más apropiado. La técnica es aplicada a un conjunto de datos reales procedentes de un estudio de captura de especies de una población. Además, se lleva a cabo un estudio de simulación para investigar su comportamiento.

Estimation of the number of cluster in a population through the generalized jackknife.

Palabras clave: Número de clusters, jacknife generalizado.

Clasificación AMS: 162G05

*Universidad Carlos III de Madrid. Dpto. de Estadística y Econometría. C/ Madrid, 126. 28903 Madrid.
– Recibido en enero de 1997.
– Aceptado en diciembre de 1997.

1. INTRODUCCIÓN

Existe una gran cantidad de trabajos en la literatura estadística sobre los métodos de estimación del número de clusters en una población, pero la mayoría de ellos han sido desarrollados entorno a la idea de que las probabilidades de observación de los distintos clusters son iguales. Ver, por ejemplo, Lewontin y Prout (1956), Darroch (1958), Harris (1968), Johnson y Kotz (1977), Darroch y Ratcliff (1982) Marchand y Schrowck (1982), Holst (1981) y Esty (1985).

Existe un concepto que está muy ligado con el de número de clusters de una población, que es el cubrimiento muestral. Se define como la suma de las probabilidades de los clusters observados en una muestra. En el caso de clusters igualmente probables, el cubrimiento viene dado por el número de clusters observados en una muestra, dividido por el número de clusters que constituyen la población. Darroch y Ratclif (1980) utilizaron exactamente la idea del cubrimiento muestral para estimar K.

Ahora bien, considerar la hipótesis de que las probabilidades de los distintos clusters son iguales es, en principio, un caso muy particular y poco frecuente. Por ejemplo, no existe una misma cantidad de animales para cada especie en un ecosistema; no se repite con la misma frecuencia cada una de las diferentes palabras que constituyen un texto; no se acuña la misma cantidad de las distintas monedas utilizadas en un país durante un centenario, etc. La mayoría de los trabajos realizados para poblaciones heterogéneas (es decir, constituidas por clusters no equiprobables) adoptan un enfoque paramétrico. Por ejemplo, Fisher, Corbet y Williams (1943) asumen que para cada cluster, el número de observaciones en la muestra se distribuye según una distribución de Poisson, y el parámetro de dicha distribución se asume que sigue una distribución Gamma. Muchos otros artículos sobre modelos de abundancia de especies en un ecosistema también hacen consideraciones paramétricas. Ver, por ejemplo, Mc Neill (1973), Engen (1978), Efron y Thisted (1976). Fue Esty (1985), el primero en estimar el número de clusters en una población heterogénea mediante el concepto de cubrimiento muestral, aunque bajo un modelo paramétrico. Chao (1992) propone una técnica de estimación no paramétrica, pero utilizando también la idea del cubrimiento muestral.

La propuesta de este trabajo es justamente plantear una técnica de estimación no paramétrica alternativa a los estimadores mencionados anteriormente, sin necesidad de plantear un modelo de probabilidad ni de recurrir al concepto de cubrimiento muestral. La técnica empleada es el método jackknife generalizado.

Por tanto, considérese una población cerrada en la cual las observaciones están agrupadas en K clusters. El significado de cerrada hace alusión a que durante el estudio no se producen entradas o salidas de clusters existentes. Se propone inicialmente un estimador sesgado, el cual es corregido y ajustado mediante el jackknife generalizado.

Este trabajo está dividido, básicamente, en tres partes. La primera presenta, fundamentalmente, la idea general del método jackknife generalizado (apartado 2). En la segunda se aplica dicho método, obteniéndose una secuencia finita de estimadores para el número de clusters en una población. Todos ellos son una combinación lineal de las frecuencias con que los clusters son observados. Se calculan sus esperanzas matematicas, sus varianzas, y además se dan intervalos de confianza (apartado 3). Para elegir el estimador más apropiado se propone un procedimiento de selección basado en contraste de hipótesis (apartado 4). En la tercera y última parte se presenta un estudio numérico realizado por simulación, para comprobar la eficacia de las principales fórmulas presentadas aquí, así como una aplicación a un conjunto de datos reales procedentes de la captura y recaptura de especies en un determinado ecosistema (apartado 5).

2. EL MÉTODO JACKKNIFE GENERALIZADO

El jackknife (Quenouille (1949)) constituye una técnica básica de remuestreo o de reutilización muestral. Quenouille introdujo el jackknife como estimador del sesgo de un estimador, y Tukey (1958) sugirió su utilidad también en la estimación de la varianza. Miller (1974) hace una revisión sobre este método y Efron (1982) presenta de una forma concisa las ideas esenciales.

Una panorámica algo diferente del jackknife es justamente el jackknife generalizado, aunque con la misma idea central de estimación del sesgo de un estimador. Ver Gray y Shucany (1972).

A continuación se presentan los pasos generales que sigue dicho método.

Sea $y_1,\dots,y_n$ una muestra aleatoria de una función de distribución $F(\theta)$. Sea $\hat{\theta}_{(n)} = \hat{\theta}(y_1,\dots,y_n)$ un estimador de θ que verifica:

$$E\left(\hat{\theta}_{(n)}\right) = \theta + (a_1/n) + \left(a_2/n^2\right) + \cdots$$

donde $a_1, a_2, \dots$ son constantes.

Sea $j_1,\dots,j_i$ una combinación de i enteros del conjunto $\{1,\dots,n\}$.

Se define $\hat{\theta}_{n-i,j_1,\dots,j_i}$ como un estimador de θ basado en $(n-i)$ observaciones, después de haber eliminado al azar $y_{j_1},\dots,j_{j_i}$ de la muestra. Entonces, el estimador

$$\hat{\theta}_{(n-i)} = \frac{\sum_{j_1<\cdots<j_i} \hat{\theta}_{n-i,j_1,\dots,j_i}}{\binom{n}{i}},$$

llamado por Fraser (1957, p. 142) el estadístico $-U$, es la base del método jackknife generalizado para la reducción del sesgo del estimador $\hat{\theta}_{(n)}$. Esto se debe porque el estimador generalizado de orden h viene dado por:

$$\hat{\theta}_{J,h} = \frac{1}{h!}\sum_{i=0}^{h} (-1)^i \binom{h}{i} (n-i)^h \hat{\theta}_{(n-i)'}$$

donde el subíndice J hace alusión al nombre de jackknife generalizado. Obsérvese que es una combinación lineal de los estimadores $\hat{\theta}_{(n-i)}$.

A continuación este método será aplicado eliminando grupos de observaciones (muestras) en vez de observaciones individuales.

3. UN ESTIMADOR PARA EL NÚMERO DE CLUSTERS EN UNA POBLACIÓN APLICANDO EL JACKKNIFE GENERALIZADO

Sea una población infinita de individuos, cerrada y formada por un número desconocido K de clusters. Sean t muestras aleatorias simples extraídas de la población con reemplazamiento, y p_j (con $j = 1,\ldots,K$) la probabilidad de observar el cluster j en cualquiera de las t muestras, $0 < p_j < 1$, $\sum_{j=1}^{K} p_j = 1$.

Sea f_r el número de clusters observados exactamente r veces de las t posibles muestras tal que,

$$S = \sum_{r=1}^{t} f_r$$

es el número de clusters observados al menos en una muestra.

Se define $\hat{K}_t = S$ un estimador natural de K obtenido a partir de la información de las t muestras, que es sesgado. Ahora el objetivo es corregir y ajustar $\hat{K}_t$ mediante su sesgo, el cual es estimado mediante el método jackknife generalizado.

Sea $\hat{K}_{t-1;r}$ un estimador de K al eliminar la muestra $r-$ésima (con $r = 1,\ldots,t$), el cual se define como:

$$\hat{K}_{t-1;r} = \hat{K}_t - \begin{pmatrix} \text{número de clusters observados} \\ \text{una vez en la muestra } r \end{pmatrix}, \quad \text{con } r = 1,\ldots,t;$$

y sea,

$$\hat{K}_{(t-1)} = \frac{1}{t}\sum_{r=1}^{t} \hat{K}_{(t-1);r}, \tag{1}$$

el estimador jackknife de primer orden, que es el promedio de los valores $\hat{K}_{t-1;r}$.

Desarrollando (1),

$$\hat{K}_{(t-1)} = \frac{1}{t}\sum_{r=1}^{t}\left\{\hat{K}_t - \left(\begin{array}{l}\text{número de clusters observados}\\ \text{una vez en la muestra } r.\end{array}\right)\right\} =$$

$$= \hat{K}_t - \frac{1}{t}\sum_{r=1}^{t}\left(\begin{array}{l}\text{número de clusters observados}\\ \text{una vez en la muestra } r.\end{array}\right) = \hat{K}_t - \frac{1}{t}\,f_1.$$

De la misma forma, si $\hat{K}_{t-2;r,1}$ es un estimador de K al eliminar las muestras r y l (con $r,l = 1,\ldots,t$) tal que,

$$\hat{K}_{t-2;r,1} = \hat{K}_t - \left\{\left(\begin{array}{l}\text{número de clusters observados}\\ \text{una vez en la muestra } r \text{ o en la } l.\end{array}\right) + \left(\begin{array}{l}\text{número de clusters observados dos}\\ \text{veces en las muestras } r \text{ y } l.\end{array}\right)\right\},$$

entonces el estimador jackknife de segundo orden es:

$$\hat{K}_{(t-2)} = \frac{1}{\binom{t}{2}}\sum_{\substack{r=1\\ l=1\\ r\neq 1}}^{t}\hat{K}_t - \left\{\left(\begin{array}{l}\text{número de clusters observados una}\\ \text{vez en la muestra } r \text{ o en la } l.\end{array}\right) -\right.$$

$$\left. - \left(\begin{array}{l}\text{número de clusters observados dos}\\ \text{veces en la muestra } r \text{ y } l.\end{array}\right)\right\} = \hat{K}_t - \frac{2}{t(t-1)}(t-1)\,f_1 - \frac{2}{t(t-1)}\,f_2 =$$

$$= \hat{K}_t - \frac{2}{t}\,f_1 - \frac{2}{t(t-1)}\,f_2.$$

$w_{m;1_1,\ldots,i_m}$ = el número de clusters observados exactamente m veces de las t posibles muestras, donde $i_1,\ldots,i_m$ es una combinación de los enteros $1,\ldots,t$.

Así, el número de clusters observados al menos una vez en las $(t-r)$ muestras es:

$$\hat{K}_{(t-r);i_1,\ldots,i_r} = \hat{K}_t - \sum_{m=1}^{r}\left\{\sum_{\{n_1,\ldots,n_m\}\subseteq\{i_1,\ldots,i_r\}} w_{m;n_1,\ldots,n_m}\right\}.$$

El sumatorio que está dentro del paréntesis es sobre las $\binom{r}{m}$ posibles combinaciones de enteros $\{n_1,\ldots,n_m\} \subseteq \{i_1,\ldots,i_r\}$. Se sigue que:

$$\hat{K}_{(t-r)} = \hat{K}_t - \frac{1}{\binom{t}{r}}\sum_{\{i_1,\ldots,i_r\}\subseteq\{1,\ldots,t\}}\sum_{m=1}^{r}\left\{\sum_{\{n_1,\ldots,n_m\}\subseteq\{i_1,\ldots,i_r\}} w_{m;n_1,\ldots,n_m}\right\} =$$

$$= \hat{K}_t - \frac{1}{\binom{t}{r}} \sum_{m=1}^{r} \binom{t-m}{r-m} \sum_{\{n_1,\ldots,n_m\} \subseteq \{1,\ldots,t\}} w_{m;n_1,\ldots,n_m} =$$

$$= \hat{K}_t - \frac{1}{\binom{t}{r}} \sum_{m=1}^{r} \binom{t-m}{r-m} \sum_{\{n_1,\ldots,n_m\} \subseteq \{1,\ldots,t\}} f_m.$$

A partir de los estimadores $\hat{K}_{(t-m)}$, el estimador jackknife generalizado para K de orden h es:

$$\hat{K}_{J,h} = \frac{1}{h!} \sum_{m=0}^{h} (-1)^m \binom{h}{m} (t-m)^h \hat{K}_{(t-m)}. \tag{2}$$

Nótese que (2) es una fórmula un poco engorrosa de manejar. A continuación se han dado valores a h desde 1 hasta 4, obteniéndose:

$$h = 1: \qquad \hat{K}_{J,1} = \hat{K}_t - \left(\frac{t-1}{t}\right) f_1,$$

$$h = 2: \qquad \hat{K}_{J,2} = \hat{K}_t + \left(\frac{2t-3}{t}\right) f_1 - \left(\frac{(t-2)^2}{t(t-1)}\right) f_2,$$

$$h = 3: \quad \hat{K}_{J,3} = \hat{K}_t + \left(\frac{3t-6}{t}\right) f_1 - \left(\frac{3t^2-15t+19}{t(t-1)}\right) f_2 + \left(\frac{(t-3)^3}{t(t-1)(t-2)}\right) f_3,$$

$$h = 4: \qquad \hat{K}_{J,4} = \hat{K}_t + \left(\frac{4t-10}{t}\right) f_1 - \left(\frac{6t^2-36t+55}{t(t-1)}\right) f_2 +$$

$$+ \left(\frac{4t^3-42t^2+148t-175}{t(t-1)(t-2)}\right) f_3 - \left(\frac{(t-4)^4}{t(t-1)(t-2)(t-3)}\right) f_4.$$

Obsérvese que $\hat{K}_{J,h}$ es una combinación lineal de las frecuencias con que los clusters son observados, de manera que dicho estimador se puede escribir de una forma general como:

$$\hat{K}_{J,h} = \sum_{r=1}^{t} a_{rh} f_r, \quad \text{con} \quad h \leqslant t,$$

donde a_{rh} son justamente los coeficientes que acompañan a las f_r los cuales están en función de t. Burnham y Overton (1978) obtuvieron este mismo resultado para la estimación del número de individuos en una población aplicando la técnica jackknife.

Justamente este artículo presenta otros resultados para los momentos del estimador $\hat{K}_{J,h}$.

Ahora, definiendo la variable aleatoria indicatriz:

$$Z_{j,r} = \begin{cases} 1 \text{ si el clusters } j \text{ es observado } r \text{ veces de las } t \text{ posibles muestras} \\ 0 \text{ en otro caso.} \end{cases}$$

$$E(\hat{K}_{J,h}) = \sum_{r=1}^{t} a_{rh}\, E(f_r) = \sum_{r=1}^{t} a_{rh}\, E\left(\sum_{j=1}^{K} Z_{j,r}\right) =$$

$$= \sum_{r=1}^{t} a_{rh} \sum_{j=1}^{K} P(Z_{j,r}=1) = \sum_{r=1}^{t} a_{rh} \sum_{j=1}^{K} \binom{t}{r} p_j^r (1-p_j)^{t-r}.$$

Haciendo $\pi_r = E(f_r) = \sum_{j=1}^{K} \binom{t}{r} p_j^r (1-p_j)^{t-r}$, tal que

$$\mathrm{Var}(f_r) = K\pi_r(1-\pi_r) \quad \text{y} \quad \mathrm{Cov}(f_r\, r_l) = -K\pi_r\pi_l,$$

se tiene que:

(3)

$$\mathrm{Var}(\hat{K}_{J,h}) = \sum_{r=1}^{t} a_{rh}^2\, \mathrm{Var}(f_r) + 2 \sum_{\substack{r=1 \\ r>1}}^{t} \sum^{t} a_{rh}\, a_{lh}\, \mathrm{Cov}(f_r\, r_l) =$$

$$= K\sum_{r=1}^{t} a_{rh}^2\, \pi_r(1-\pi_r) - 2K \sum_{\substack{r=1 \\ r>1}}^{t} \sum^{t} a_{rh}\, a_{lh}\, \pi_r\, \pi_l =$$

$$= K\sum_{r=1}^{t} a_{rh}^2\, \pi_r - K\sum_{r=1}^{t} a_{rh}^2\, \pi_r^2 - 2K \sum_{\substack{r=1 \\ r>1}}^{t} \sum^{t} a_{rh}\, a_{lh}\, \pi_r\, \pi_l =$$

$$= K\sum_{r=1}^{t} a_{rh}^2\, \pi_r - K\sum_{r=1}^{t}\sum_{l=1}^{k} a_{rh}\, \pi_r\, a_{lh}\, \pi_l = K\sum_{r=1}^{t} a_{rh}^2\, \pi_r - K\sum_{r=1}^{t} a_{rh}\, \pi_r \sum_{l=1}^{t} a_{lh}\, \pi_l =$$

$$= K\sum_{r=1}^{t} a_{rh}^2\, \pi_r - \frac{K\sum_{r=1}^{t} a_{rh}\, \pi_r\, K\sum_{l=1}^{t} a_{lh}\, \pi_l}{K} = K\sum_{r=1}^{t} a_{rh}^2\, \pi_r - \frac{E(\hat{K}_{J,h})\, E(\hat{K}_{J,h})}{K} =$$

$$= K\sum_{r=1}^{t} a_{rh}^2\, \pi_r - \frac{E^2(\hat{K}_{J,h})}{K}.$$

Un estimador insesgado de mínima varianza para $\mathrm{Var}(\hat{K}_{J,h})$ es:

$$\widehat{\mathrm{Var}}(\hat{K}_{J,h}) = \frac{\hat{K}_t}{\hat{K}_t - 1}\left\{\hat{K}_t \sum_{r=1}^{t} a_{rh}^2\ \hat{f}_r - \frac{(\hat{K}_{J,h})^2}{\hat{K}_t}\right\}, \tag{4}$$

donde $\hat{f}_r$ son los valores observados de π_r.

Por otra parte, $(f_1, f_2, \ldots, f_t)$ es una variable aleatoria multinomial tal, que cualquier combinación lineal de éstas es bien sabido que se distribuye asintóticamente como una normal. Así $\hat{K}_{J,h}$ se distribuye aproximadamente como:

$$N\left(E(\hat{K}_{J,h}),\ \widehat{\mathrm{Var}}(\hat{K}_{J,h})\right).$$

Sea $z_{\alpha/2}$ el percentil $1-(\alpha/2)$ de la distribución $N(0,1)$. Un intervalo de confianza para K con un nivel de confianza $1-\alpha$ es:

$$\left(\hat{K}_{j,h} - z_{\alpha/2}\left(\widehat{\mathrm{Var}}(\hat{K}_{J,h})\right)^{1/2},\ \hat{K}_{J,h} + z_{\alpha/2}\left(\widehat{\mathrm{Var}}(\hat{K}_{J,h})\right)^{1/2}\right).$$

4. UN PROCEDIMIENTO DE ESTIMACIÓN

Una vez calculados los estimadores $\hat{K}_{J,h}$ para distintos valores de h, cabe preguntarse cuál es el que mejor estima o menor error comete. Para ello se propone el siguiente procedimiento basado en contrastes de hipótesis.

Contrastar la hipótesis nula:

$$H_{0,h} : E(\hat{K}_{J,h+1} - \hat{K}_{J,h}) = 0$$

frente a la hipótesis alternativa:

$$H_{1,h} : E(K_{J,h+1} - K_{J,h}) \neq 0,$$

y elegir como estimador de K, $\hat{K} = \hat{K}_{J,h}$, tal que $H_{0,h}$ es la primera hipótesis nula no rechazada. El mecanismo es el siguiente. Si $h = 1$ y H_{01} no es rechazada es que hay evidencia de que se logre un descenso en el sesgo por utilizar $\hat{K}_{J,2}$, mucho más que utilizando $\hat{K}_{J,1}$ y, por consiguiente, se concluye que no hay razón para utilizar $\hat{K}_{J,2}$. De hecho debe utilizarse $\hat{K}_{J,1}$ como estimador de K debido a que posteriores estimadores con $h > 2$ tendrán un sesgo mucho mayor. Si H_{01} es rechazada, entonces cabe esperar una reducción del sesgo utilizando $\hat{K}_{J,2}$. Ante la aceptación de $\hat{K}_{J,2}$ como estimador de k se contrastaría la elección de $\hat{K}_{J,2}$ frente a la de $\hat{K}_{J,3}$. En caso de rechazo de H_{02}

se volvería a realizar el contraste correspondiente. El proceso continua de manera progresiva hasta conseguir el estimador secuencial, $\hat{K}_{J,h}$, cuya H_{0h} sea la primera no rechazada.

Nótese que bajo la condición de que $H_{0,h}$ sea verdadera, el estadístico del contraste es:

$$T_h = \frac{\hat{K}_{J,h+1} - \hat{K}_{J,h}}{\left(\widehat{\mathrm{Var}}(\hat{K}_{J,h+1} - \hat{K}_{J,h}/\hat{K}_t)\right)^{1/2}},$$

el cual se distribuye aproximadamente como una $N(0,1)$. Para un nivel de significación α, la región crítica es:

$$C = \left\{z : |z| > z_{\alpha/2}\right\},$$

siendo $z_{\alpha/2}$ el valor de una normal (0,1) que deja a la derecha un área de probabilidad igual a $\alpha/2$. Por consiguiente se rechaza $H_{0,h}$ si $T_\alpha > z_{\alpha/2}$.

Es necesario indicar que como $\hat{K}_{J,h} = \sum_{r=1}^{t} a_{rh}\ f_r$, entonces

$$\hat{K}_{J,h+1} - \hat{K}_{J,h} = \sum_{r=1}^{t} a_{rh}\ f_r,$$

es decir, es también una combinación lineal de las frecuencias f_r. Por tanto, se sigue de la expresión (4):

$$\widehat{\mathrm{Var}}(\hat{K}_{J,h+1} - \hat{K}_{J,h}) = \frac{\hat{K}_t}{\hat{K}_t - 1}\left\{\hat{K}_t \sum_{r=1}^{t} a_{rh}^2\ \hat{f}_r - \frac{(\hat{K}_{J,h+1} - \hat{K}_{J,h})^2}{\hat{K}_t}\right\}. \tag{5}$$

5. ESTUDIOS NUMÉRICOS

Ejemplo 1

La evaluación de los estimadores $\hat{K}_{J,h}$ ha sido llevada a cabo simulando t muestras (en particular 5, 10 y 15) de tamaño 100 de una población de K clusters (se han considerado desde una población con pocos clusters, $K = 5$, hasta una población constituida por numerosos clusters, $K = 150$). Cada caso se ha simulado 100 veces, de manera que los datos presentados son promedio de los resultados. También se ha calculado la varianza y desviación típica de los $\hat{K}_{J,h}$ para los casos de mayor interés, $K = 150$ (con $t = 5$ y $t = 15$). Nótese que son justamente los casos donde la estima

de K puede tener más sesgo debido a la heterogeneidad de la población. Obsérvese que de la expresión (5) se obtiene:

$$\widehat{\mathrm{Var}}(\hat{K}_{J,1}) = \hat{K}_t \left(1+\frac{t-1}{t}\right)^2 f_1 + (\hat{K}_t - f_1) - \hat{K}_{J,1},$$

$$\widehat{\mathrm{Var}}(\hat{K}_{J,2}) = \hat{K}_t \left(1+\frac{2t-3}{t}\right)^2 f_1 + \hat{K}_t \left(1-\frac{(t-2)^2}{t(t-1)}\right)^2 f_2 + \left(\hat{K}_t - (f_1+f_2)\right) - \hat{K}_{J,2},$$

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{K}_{J,3}) = {} & \hat{K}_t \left(1+\frac{3t-6}{t}\right)^2 f_1 + \hat{K}_t \left(1-\frac{3t^2-15t+19}{t(t-1)}\right)^2 f_2 + \\ & + \hat{K}_t \left(1+\frac{(t-3)^2}{t(t-1)(t-2)}\right)^2 f_3 + \left(\hat{K}_t - (f_1+f_2+f_3)\right) - \hat{K}_{J,3}\end{aligned}$$

y

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{K}_{J,4}) = {} & \hat{K}_t \left(1+\frac{4t-10}{t}\right)^2 f_1 + \hat{K}_t \left(1-\frac{6t^2-36t+55}{t(t-1)}\right)^2 f_2 + \\ & + \hat{K}_t \left(1+\frac{4t^3-42t+148t-175}{t(t-1)(t-2)}\right)^2 f_3 + \hat{K}_t \left(1-\frac{(t-4)^4}{t(t-1)(t-2)(t-3)}\right)^2 f_4 + \\ & + \left(\hat{K}_t - (f_1+f_2+f_3+f_4)\right) - \hat{K}_{J,4}.\end{aligned}$$

De los resultados obtenidos y reflejados en las tablas 1 y 2 se deduce que:

- $\hat{K}_j$ es sesgado, infraestimando siempre K, es decir, estima por debajo del valor de K.
- Para poblaciones constituidas por pocos clusters ($K < 50$) las estimas de $\hat{K}_{J,h}$ (con $h = 1,2,3,4$) son excelentes. Obsérvese que justamente para $K = 50$, los valores de $\hat{K}_{J,h}$ (con $j = 1,2,3$) empiezan a ser estimas sesgadas.
- En poblaciones constituidas por un gran número de clusters ($K = 100$ o $K = 150$), los estimadores $\hat{K}_{J,h}$ con $j = 1,\ldots,4$, trabajan muy bien, no produciéndose un sesgo muy grande. Nótese que a medida que t crece, la diferencia entre K y $\hat{K}_{J,h}$ es cada vez menor.
- En cualquiera de los casos, el mayor sesgo se produce para el estimador $\hat{K}_{J,1}$. Obsérvese siempre la mejora de $\hat{K}_{J,4}$ con respecto a $\hat{K}_{J,1}$.
- A medida que el número de muestras extraídas es mayor, la estimación en general de K con esta técnica es mejor.

- Si K es pequeño, el número de muestras en que cada cluster es observado, de las t posibles, es muy elevado. Si K es grande, raramente existen clusters que son observados en la mayoría de las t muestras.
- En la tabla 2 se observa que la desviación típica es menor a medida que el número de muestras crece. Además, en cualquiera de los tres casos de interés, cuando h crece, la desviación típica decrece.

Tabla 1. *Cálculo de los estimadores* $\hat{K}_{J,h}$

Caso	k	t	f_r	$\hat{K}_{J,h}$		
1	150	5	44, 42, 23, 5, 0	$\hat{K}_J = 114$	$\hat{K}_{J,1} = 159.20$	$\hat{K}_{J,2} = 156.36$
				$\hat{K}_{J,3} = 153.96$	$\hat{K}_{J,4} = 152.86$	
2		10	20, 26, 36, 33, 9	$\hat{K}_J = 134$	$\hat{K}_{J,1} = 152.0$	$\hat{K}_{J,2} = 149.51$
			9, 1, 0, 0, 0	$\hat{K}_{J,3} = 152.22$	$\hat{K}_{J,4} = 151.11$	
3		15	4, 6, 35, 27, 34	$\hat{K}_J = 150$	$\hat{K}_{J,1} = 153.37$	$\hat{K}_{J,2} = 152.37$
			22, 7, 8, 4, 2	$\hat{K}_{J,3} = 152.05$	$\hat{K}_{J,4} = 151.75$	
			1, 0, 0, 0, 0			
4	100	5	34, 45, 9, 1, 0, 0	$\hat{K}_J = 89$	$\hat{K}_{J,1} = 116.20$	$\hat{K}_{J,2} = 116.35$
				$\hat{K}_{J,3} = 97.98$	$\hat{K}_{J,4} = 98.97$	
5		10	9, 6, 32, 22, 12	$\hat{K}_J = 88$	$\hat{K}_{J,1} = 106.10$	$\hat{K}_{J,2} = 101.92$
			5, 2, 0, 0, 0	$\hat{K}_{J,3} = 102.65$	$\hat{K}_{J,4} = 98.83$	
6		15	1, 5, 8, 20, 26	$\hat{K}_J = 90$	$\hat{K}_{J,1} = 90.93$	$\hat{K}_{J,2} = 87.77$
			12, 5, 9, 2, 0	$\hat{K}_{J,3} = 98.82$	$\hat{K}_{J,4} = 100.96$	
			2, 0, 0, 0, 0			
7	50	5	12, 6, 5, 2, 0	$\hat{K}_J = 25$	$\hat{K}_{J,1} = 43.62$	$\hat{K}_{J,2} = 48.60$
				$\hat{K}_{J,3} = 49.95$	$\hat{K}_{J,4} = 49.93$	
8		10	2, 4, 11, 5, 7	$\hat{K}_J = 33$	$\hat{K}_{J,1} = 46$	$\hat{K}_{J,2} = 49.87$
			2, 2, 0, 0, 0	$\hat{K}_{J,3} = 50.73$	$\hat{K}_{J,4} = 50.87$	
9		15	0, 1, 4, 8, 8	$\hat{K}_J = 42$	$\hat{K}_{J,1} = 48$	$\hat{K}_{J,2} = 49.89$
			7, 9, 3, 0, 2	$\hat{K}_{J,3} = 50.29$	$\hat{K}_{J,4} = 50.42$	
			0, 0, 0, 0, 0			
10	20	5	1, 1, 4, 9, 2	$\hat{K}_J = 17$	$\hat{K}_{J,1} = 21.80$	$\hat{K}_{J,2} = 19.95$
				$\hat{K}_{J,3} = 20.98$	$\hat{K}_{J,4} = 20.35$	

Continúa

Continúa

Caso	k	t	f_r	$\hat{K}_{J,h}$		
11	20	10	0, 1, 0, 1, 1 3, 2, 6, 4, 0	$\hat{K}_J = 18.00$ $\hat{K}_{J,3} = 20.86$	$\hat{K}_{J,1} = 19.35$ $\hat{K}_{J,4} = 20.36$	$\hat{K}_{J,2} = 21.08$
12		15	0, 0, 1, 0, 0 0, 0, 2, 3, 2 3, 5, 3, 0, 0	$\hat{K}_J = 19.00$ $\hat{K}_{J,3} = 20.63$	$\hat{K}_{J,1} = 20.63$ $\hat{K}_{J,4} = 20.23$	$\hat{K}_{J,2} = 19.45$
13	10	5	0, 0, 1, 2, 7	$\hat{K}_J = 10.00$ $\hat{K}_{J,3} = 10.65$	$\hat{K}_{J,1} = 10.75$ $\hat{K}_{J,4} = 10.45$	$\hat{K}_{J,2} = 10.29$
14		10	0, 0, 0, 0, 0 0, 1, 1, 3, 3	$\hat{K}_J = 8.00$ $\hat{K}_{J,3} = 10.24$	$\hat{K}_{J,1} = 9.27$ $\hat{K}_{J,4} = 10.19$	$\hat{K}_{J,2} = 9.61$
15		15	0, 0, 0, 0, 0 0, 0, 0, 0, 0 1, 1, 3, 3, 1	$\hat{K}_J = 9.00$ $\hat{K}_{J,3} = 9.94$	$\hat{K}_{J,1} = 9.73$ $\hat{K}_{J,4} = 10.03$	$\hat{K}_{J,2} = 10.47$
16	5	5	0, 0, 0, 0, 5	$\hat{K}_J = 4.86$ $\hat{K}_{J,3} = 5.00$	$\hat{K}_{J,1} = 4.90$ $\hat{K}_{J,4} = 5.00$	$\hat{K}_{J,2} = 5.00$
17		10	0, 0, 0, 0, 0 0, 0, 0, 0, 5	$\hat{K}_J = 4.97$ $\hat{K}_{J,3} = 5.00$	$\hat{K}_{J,1} = 5.09$ $\hat{K}_{J,4} = 5.00$	$\hat{K}_{J,2} = 5.00$
18		15	0, 0, 0, 0, 0 0, 0, 0, 0, 0 0, 0, 0, 0, 5	$\hat{K}_J = 4.94$ $\hat{K}_{J,3} = 5.00$	$\hat{K}_{J,1} = 5.12$ $\hat{K}_{J,4} = 5.00$	$\hat{K}_{J,2} = 5.00$

Tabla 2. *Cálculo de la varianza de los* $\hat{K}_{J,h}$ *para los casos 1 y 2*

t	h	$\widehat{\mathrm{Var}}(\hat{K}_{J,h})$	$\left(\widehat{\mathrm{Var}}(\hat{K}_{J,h})\right)^{1/2}$
5	1	80.82	8.99
	2	52.78	7.26
	3	9.92	3.15
	4	8.76	2.96
10	1	8.00	2.83
	2	2.40	1.55
	3	0.16	0.41
	4	0.11	0.34
15	1	16.81	4.10
	2	11.22	3.35
	3	1.44	1.20
	4	1.06	1.03

Nótese que los valores de las f_r y K_J son valores promedio de las 100 simulaciones realizadas.

A continuación se procede a la elección del estimador secuencial para los casos 2 y 3. Para ello se aplicará el siguiente algoritmo de contraste de hipótesis explicado anteriormente.

Algoritmo

Paso 1: Hacer $h = 1$. Contrastar la hipótesis nula

$$H_{0,h} : E(\hat{K}_{J,h+1} - \hat{K}_{J,h}) = 0$$

frente a la hipótesis alternativa

$$H_{1,h} : E(\hat{K}_{J,h+1} - \hat{K}_{J,h}) \neq 0.$$

Paso 2: Rechazar $H_{0,h}$ si $|T_\alpha| > z_{\alpha/2}$, con $z_{\alpha/2}$ el valor de una $N(0,1)$ que deja a ambos lados un área de probabilidad $\alpha/2$.
Ir al paso 3.
En caso contrario, $h = h+1$ e ir al paso 1.

Paso 3: Un estimador para K es $\hat{K}_{J,h}$.

Entonces, para el caso 3 ($K = 150, t = 15$), en el primer contraste se tiene que:

$$T_1 = \frac{152.37 - 153.73}{\left(\frac{150}{149}\left(6.89 - \frac{1.84}{150}\right)\right)^{1/2}} = 0.51.$$

Como $|T_1| < z_{\alpha/2}$ (con $\alpha = 0.05$, $z_{0.025} = 1.96$), se acepta la hipótesis nula y, por consiguiente, hay que realizar el siguiente contraste.

El estadístico del segundo contraste es $T_2 = 3.22$. Como $|T_2| > z_{\alpha/2}$ (con $\alpha = 0.05$), se rechaza la hipótesis nula, siendo el estimador apropiado para K, $\hat{K} = \hat{K}_{J,2} = 152.37$.

Los valores de T_h (para $h = 1,2,3,4$) del caso $K = 150$ ($t = 10$) están dispuestos en la tabla 3. Dependiendo del nivel de significación que se fije se obtendrá un estimador apropiado para K. Por ejemplo, para $\alpha = 0.05$, $z_{0.025} = 1.96$, $|T_2| = 0.11 < z_{0.025}$, por tanto se acepta la hipótesis nula. En cambio, $|T_3| = 3.51 > z_{0.025}$ y, por consiguiente, el estimador apropiado es $\hat{K}_{J,3} = 150.32$.

Tabla 3. *Cálculo de los T_h para el caso 2 de la tabla 1*

h	$\hat{K}_{J,h}$	T_h
1	152.00	−0.48
2	149.00	0.11
3	150.32	3.51
4	155.34	4.55

Un intervalo de confianza para los casos 2 y 3 son respectivamente:

$$\textit{Caso 2:}\ (150.32 \pm 1.96 \times 0.41) = (149.51,\ 151.12)$$

$$\textit{Caso 3:}\ (152.37 \pm 1.96 \times 3.35) = (145.81,\ 158.93)$$

Ejemplo 2

Como aplicación del procedimiento desarrollado en el apartado 3, se ha considerado un conjunto de datos reales pertenecientes a 10 muestras extraídas de un pantano junto al río Pettaquamscutt al sur de la Isla de Rhode. Son datos recogidos en abril de 1978 por Jeffrey Hyland del Colegio de Oceanografía de la Universidad de la Isla de Rhode. En Heltshe y Forrester (1983) se utilizan estos datos para realizar un estudio de muestreo, indicando que el número de clases de especies existentes en el pantano es 21. En las 10 muestras se observaron 14 especies cuyos datos fueron:

Tabla 4

	Muestras									
Especies observadas	1	2	3	4	5	6	7	8	9	10
S. benedicti.		13	21	14	5	22	13	4	4	27
N. succines.	2	2	4	1	1	1	1		1	6
P. lignis.		1						1		
S. Robustus.	1		1	2		6	1		1	2
E. heteropoda.			1	2						1
H. filiformis.	1	1	2	1		1			1	5
C. capitata.	1									
S. viridis.	2									
H. grayi.		1								
B. clavata.			1							
M. balthica.			3							2
A. abdita.			5	1		2				3
N. texana.								1		
Tubifocodies sp.	8	36	14	19	3	22	6	8	5	41

La tabla 3 indica por filas el número de observaciones realizadas por cada especie en cada una de las muestras. Se deduce fácilmente que:

$$f_1 = 5; \quad f_2 = 2; \quad f_3 = 0; \quad f_4 = 2; \quad f_5 = 0,$$
$$f_6 = 1; \quad f_7 = 1; \quad f_8 = 0; \quad f_9 = 2; \quad f_{10} = 1.$$

Entonces los valores de los estimadores son:

$$\hat{K}_{J,1} = 19.9; \quad \hat{K}_{J,2} = 19.9; \quad \hat{K}_{J,3} = 20.01; \quad \hat{K}_{J,4} = 20.01$$

los cuales se aproximan al verdadero valor de K.

6. AGRADECIMIENTOS

Quiero dar las gracias al profesor Kenneth P. Burnham, profesor de Estadística del Departamento de Biología de la Universidad de Colorado, y al profesor Kenneth H. Pollock del Departamento de Estadística de la Universidad de Carolina, por la colaboración prestada para la realización de este artículo.

7. BIBLIOGRAFÍA

[1] **Burnham, K.P.** y **Overton, W.S.** (1978). «Estimation of the size of a closed population when capture probabilities vary among animals». *Biometrika*, **63, 3**, 625–633.

[2] **Chao, A.** (1992). «Estimating the number of classes via sample coverage». *Journal of the American Association*, **87, 417**, 210–217.

[3] **Darroch, J.N.** (1958). «The multiple recapture census I: Estimation of a closed population». *Biometrika*, **40**, 343–359.

[4] **Darroch, J.N.** y **Ratclif, D.** (1980). «A note on capture-recapture estimation». *Biometrika*, **45**, 343–359.

[5] **Efron, B.** (1979). «Bootstrap methods, Another look at the jackknife». *The Annals of Statistic*, **7**, 1–26.

[6] **Efron, B.** y **Thisted, R.** (1976). «Estimating the number of unseen species: How many words did Shakespeare Know?». *Biometrics*, **63**, 435–447.

[7] **Efron, B.** (1982). «The jackknife, the bootstrap and other resampling plans». *Regional conference series and applied mathematics*. CBMS-NSF.

[8] **Engen, S.** (1978). *Stochastic Abundance Models*. London: Chapman-Hall.

[9] **Esty, W.W.** (1985). «The estimation of the number os classes in a population and the coverage of a sample». *Mathematical Scientist*, **10**, 41–50.

[10] **Fraser, D.A.S.** (1957). *Nonparametric methods in Statistic*. New York: Wiley.

[11] **Fiher, R.A., Corbet, A.S.** y **Williams, C.B.** (1943). «The relation between the number of species and the number of individuals in a random sample of an animal population». *Journal of Animal Ecology*, **12**, 42–58.

[12] **Gray, H.L.** y **Shucany, W.R.** (1972). *The Generalized Jackknife Statistic*, New York: Marcel Dekker.

[13] **Harris, B.** (1968). «Statistical inference in the classical occupancy problem unbiased estimation of the number of classes». *Journal of the American Statistical Association*, **63**, 837–847.

[14] **Heltshe, J.F.** y **Forrester, N.E.** (1983). «Estimating Species Richness using the Jackknife Procedure». *Biometrics*, **39**, 1–11.

[15] **Holst, L.** (1981). «Some asymptotic result for incomplete multinomial or Poisson samples». *Scandinavian Journal of Statistic*, **8**, 243–246.

[16] **Johnson, N.L.** y **Kotz, S.** (1977). *Urn models and their applications: An approach to modern discrete probability theory*. New York: John Wiley.

[17] **Lewontin, R.C.** y **Prout, T.** (1956). «Estimation of the number of different classes in a population». *Biometrics*, **12**, 211–223.

[18] **Marchand, J.P.** y **Schroeck, P.E.** (1982). «On the estimation of the number of equally likely classes in a population». *Communications in Statistics, Part A-Theory and Methods*, **11**, 1139–1146.

[19] **McNeil, D.** (1973). «Estimating an Author's vocabulary». *Journal of the American Statistical Association*, **68, 341**, 92–97.

[20] **Miller, R.G.** (1974). «The Jackknife-a rewiew». *Biometrika*, **61**, 1–17.

[21] **Quenouille, M.H.** (1949). «Approximate tests of correlation in time series». *Journal of the Royal Statistical Society, Series B*, **11**, 68–84.

[22] **Tukey, J.** (1958). «Bias and confidence in not quite large samples abstract». *Annals of Mathematical Statistics*, **29**, 614.

ENGLISH SUMMARY

ESTIMATION OF THE NUMBER OF CLUSTERS IN A POPULATION THROUGH THE GENERALIZED JACKKNIFE

J.J. PRIETO MARTÍNEZ
Universidad Carlos III de Madrid*

Let a population be compared by an unknown number of clusters. This work develops a finite sequence of estimators for the number of clusters using the generalized jackknife method. It is found that there estimators are a linear combination of the observation frequencies. A procedure is then proposed for selecting just one of them which will be finaly used. The technique is applied to an real set of capture histories from a study of a population. Also the paper includes a simulation study toinvestigate its behavior.

Keywords: Number of clusters, generalized jackknife.

AMS Classification: 162G05

* Universidad Carlos III de Madrid. Dpto. de Estadística y Econometría. C/ Madrid, 126. 28903 Madrid.
– Received January 1997.
– Accepted December 1997.

Assume that t random samples of size n are drawn from a population with unknown number of clusters, K, and unequal cluster probabilities.

There is a large statistical literature on estimation methods of the number of clusters in a population, In most practical applications, the equally likely assumption is not valid. For practical examples, see Lewontin and Prout (1956), Darroch (1958), Harris (1968), Johnson and Kotz (1977), Darroch and Ratclif (1982), Marchand and Schrowch (1982), Holst (1981) and Esty (1985).

The sample coverage, C, of a random sample from a multinomial population is defined to be the sum of the probabilities of the observed clusters. For an equiprobable population, C reduces to D/K where D is the number of distinct clusters observed in the sample. Darroch and Ratclif (1980) exactly used the idea of sample coverage to estimate K.

Most authors adopted a parametric approach to handle heterogenenous populations (i.e., unequal clusters probabilities). For example, Fisher, Corbet and Williams (1943) assumed that for each cluster the number of elements observed in the sample follows a Poisson distribution and the Poisson parameter is assumed to have a gamma-type distribution. Many other papers on stochastic abundance models also make parametric assumption; see, for example, McNeill (1973), Efron and Thisted (1976) and Engen (1978). For heterogeneous populations, Esty (1985) was the first to apply the concept of sample coverage to estimate the number of clusters in a parametric setup. Chao (1992) proposes a nonparametric estimation method but using also the idea the coverage sample.

In this paper, a nonparametric estimation technique is developed based on the generalized jackknife. Note that this method does't assume a parametric model and it doesn't use the sample coverage.

Then, let t random samples are taken from a population of elements belonging to K different clusters. Let p_j be the probability that any observation belongs to the $j-$th cluster in some of the t samples, with $0 < p < 1$, $\sum p_j = 1$.

Let f_r the number of clusters observed exactly r times of the t possibles samples, and

$$S = \sum_{r=1}^{t} f_r$$

the number of clusters observed.

The $\hat{K}_t = S$ is a natural estimator to K. The goal is correcting and adjusting for the bias $\hat{K}_t$; the approach to estimating the bias is under the method generalized jackknife.

The estimator obtained is given by

$$\hat{K}_{J,h} = \sum_{r=1}^{t} a_{rh}\ f_r,$$

which is a linear combination of the frequencies whose clusters are observed.

The expected value is calculated:

$$E(\hat{K}_{J,h}) = \sum_{r=1}^{t} a_{rh} \sum_{j=1}^{K} \binom{t}{r}\ p_j^r (1-p_j)^{t-r};$$

and an estimator to the variance is:

$$\widehat{\mathrm{Var}}(\hat{K}_{J,h}) = \frac{\hat{K}_t}{\hat{K}_t - 1} \left\{ \hat{K}_t \sum_{r=1}^{t} a_{rh}^2\ \hat{f}_r - \frac{(\hat{K}_{J,h})^2}{\hat{K}_t} \right\}.$$